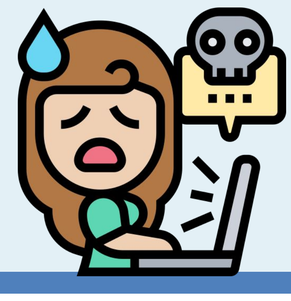
BiRDy: Bullying Role Detection in Multi-Party Chats

Anaïs Ollagnier[1], Elena Cabrio[1], Serena Villata[1], Sara Tonelli[2]

[1]Université Côte d'Azur, Inria, CNRS, I3S

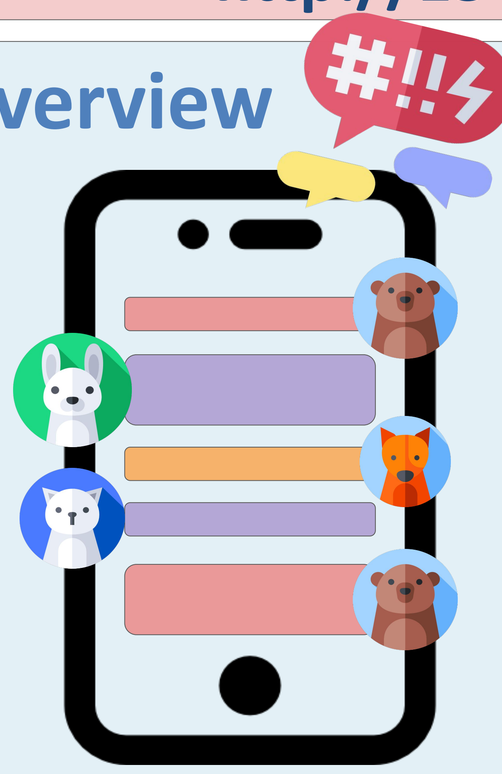
[2]Fondazione Bruno Kessler, Trento, Italy



BiRDy is a solution aiming at <u>automating the identification of the different participant role of involvement in cyberbullying episodes</u>, e.g. victim, bully, conciliator, bystanders assistant or defender

http://134.59.134.227/demo_prd/index.html

BirDy Overview



input:
multi-party chats

pipeline:
participant role classification
role scoring function

output:
unique role assignment

PRI classification

<u>Preprocess</u>:
Exploring the use of **contextual window** by collecting for each message the ***n* previous sentences authored by the same user;**

<u>Modelling</u>:

Classifying exchanged messages according to the role of the author (e.g. victim, bully, bystanders, conciliator) using **fine-tuned multilingual BERT models**.

Role scoring

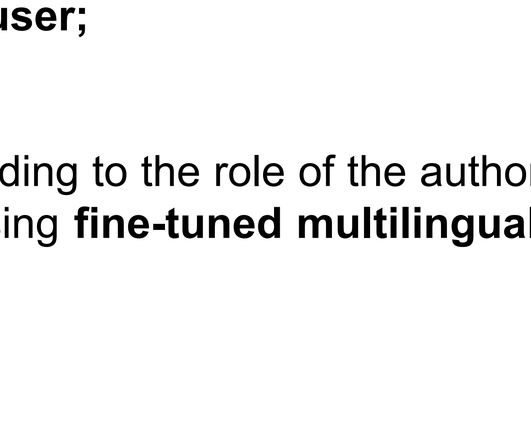
Assign a unique role to users by considering the probabilities established by the model.

$$Role\_scoring(u,l) = \frac{\sum_{\substack{m \in speech\_acts(u) \\ argmax(D(m))=l}} P_{am}(m,l)}{|speech\_acts(u)|}$$

Confidence Score

Detailing for each label the **degree of likelihood** per user.

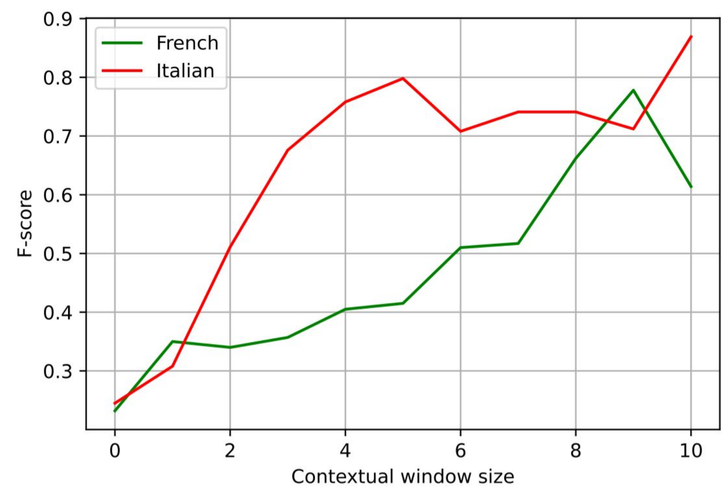
$$Confidence(u,l) = \frac{|U| \times Role\_scoring(u,l)}{\sum_{u' \in U} Role\_scoring(u',l)}$$



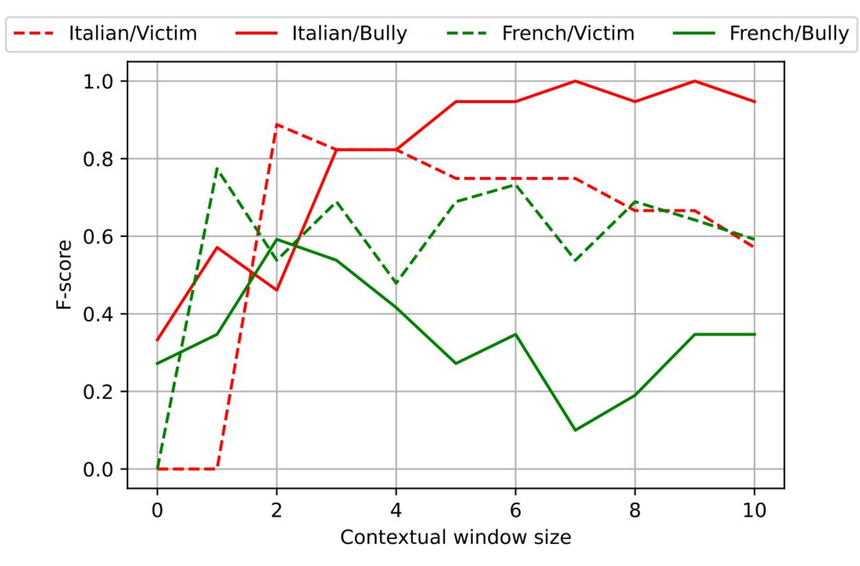
Evaluation

<u>Datasets</u>:
French: aollagnier/CyberAgressionAdo-v1
Italian: dhfbk/WhatsApp-Dataset

<u>PRI classification</u>



<u>Bully/Victim classification</u>



Acknowledgements

This work is supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002 and the EFELIA Côte d'Azur project ANR-22-CMAS-0004 .